International Journal of Computer and Electrical Engineering, Vol. 1, No. 5 December, 2009
1793-8163

Improving Machine Translation using Hybrid Dictionary-Graph Based Word Sense Disambiguation with Semantic and Statistical Methods

Ola Mohammad Ali, Mahmoud GadAlla, and Mohammad Said Abdelwahab

Abstract— **This paper describes a machine translation prototype in which noun phrase translation is defined as a subtask of machine translation. A dedicated noun phrase translation subsystem is built and improved to translate Arabic noun phrase into English using only minimal resources for both the source and the target language. This work proposes a dictionary-graph based WSD approach to improve machine translation using hybrid semantic-statistical method based on computing words relatedness and a statistical measure of association to get the relation between ambiguous words. This relation was used with viterbi search algorithm to find the appropriate translation of the Arabic noun phrase. A shallow source language analysis, combined with a translation dictionary and a mapping system of source language phenomena into the target language and a target language corpus for generation are all the resources needed in the described system.**

***Index Terms*— Arabic Noun Phrase, Limited Resources, Machine Translation, Monolingual Corpus, Word Sense Disambiguation.**

I. Introduction

In current Natural Language Processing (NLP) technology, however, machine translation relies heavily on expensive resources, such as large parallel corpora and expensive tools such as parsers and semantic taggers. Consequently, the number of languages that have such advanced technology at their disposal is small. While machine translation industrial technologies are mainly rule-based, current research is mainly on data-driven methods such as Statistical Machine Translation (SMT). Most SMT stems rely on parallel corpora, and the development of a Rule Based Machine Translation (RBMT) system is a tedious and very expensive undertaking. This system looking for a low-cost solution, so it did not consider pure RBMT. Both purely data-driven and purely rule-based approaches each have their intrinsic obstacles. [12], [15], [20] suggested that a hybrid approach is the way to go.

RBMT is only expensive if you try to model fine grained distinctions. Taggers and shallow rule-based parsers are relatively easy to obtain. Similarly many SMT approaches are hard tasks since sufficient parallel material is needed to model the whole translation process. On the other hand, more and more monolingual corpora of reasonable size are becoming available for an ever-increasing set of languages. Therefore, investigation of machine translation with limited resources is receiving more and more attention.

This paper presents a rule-based statistical machine translation system which translate Arabic noun phrase into English. Rule-based methods are used where representations and decisions can be determined a-priori with high accuracy based on linguistic insight. Corpora serve as a basis to ground decisions where uncertainty remains. SMT methods are used for target language generation, using only a target language corpus and a bilingual dictionary instead of a parallel corpus. Translating Noun Phrase (NP) is a very important task toward sentence translation since NPs form the majority of textual content of the scientific and technical documents.

The rest of the paper is structured as follows. In, Section II, the structure of the system is presented. It also describes the proposed disambiguation approach as a part of target language generation module. In Section III, experiments and results is described. Section IV, gives some concluding remarks and future directions.

II. Structure Of The System

The main goal of this system is to build a translation system without parallel corpora and without an extensive rule-set. Fig. 1 shows the architecture of the Arabic to English MT system and which resources are used at which stage in the translation process. The system consists of three components: Source Language Analysis, Source to Target Transfer, and Target Language Generation.

Manuscript received May 15, 2009.

O. M. Ali, is with Department of Scientific Computing, Faculty of Computer and Information Systems, Ain Shams University, Cairo, Egypt (fax:+2-02-26849677;e-mail: olaforn@yahoo.com)

M. GadAlla, was with Military Technical Collage, Cairo, Egypt. He is now with the Department of Computer Science, Modern Academy, Maadi, Cairo, Egypt (e-mail: mahmoud_mtc@yahoo.com).

M. S. Abdelwahab is with Department of Scientific Computing, Faculty of Computer and Information Systems, Ain Shams University, Cairo, Egypt (e-mail: m_s_wahab@fcisainshams.edu.eg)

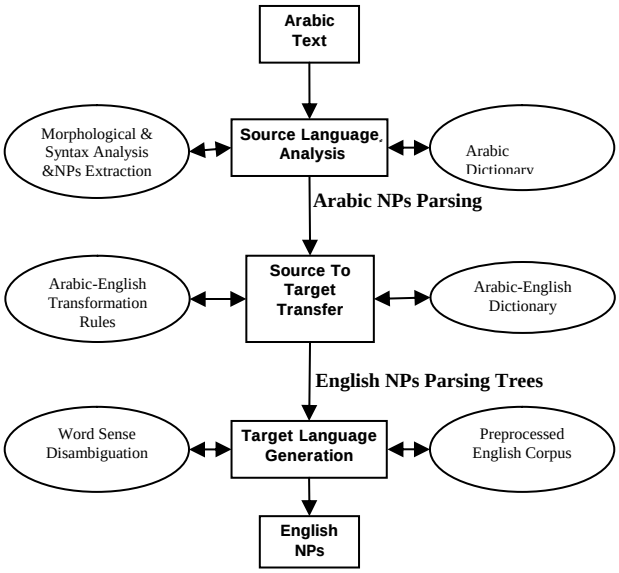

Fig. 1 The architecture of the Arabic to English MT system

There are a number of translation problems, which need to be solved. The most difficult problems in developing translation systems for Arabic are the predominance of, the morphological complexity and non-diacritized text material.

Morphological Complexity: Arabic has a rich and productive morphology which leads to a large number of potential word forms. This increases the out-of-vocabulary rate and prevents the robust estimation of language model probabilities.

Script Representation: The Arabic alphabet only contains letters for long vowels and consonants. Short vowels and other pronunciation phenomena, like consonant doubling, can be indicated by diacritics (short strokes placed above or below the preceding consonant). However, Arabic texts are almost never fully diacritized and are thus potentially unsuitable for recognizer training, this leads to many problems. The absence of this information leads to many identical-looking word forms (e.g. the form كتب (ktb) can correspond to *kataba* which means *he wrote*, *kutub*, which means *books* or 19 other forms) in a large variety of contexts, which decreases predictability in the language model.

These translation problems are tackled at different places within the translation system architecture, corresponding to the sections in this paper: the Arabic morphological analyzer and parser (*A*) and search engine(*C*). Morphological complexity problems are solved using the Arabic morphological analyzer and parser. Script representation problems are solved using both Arabic-English dictionary and searching engine. While the dictionary would generate several meaning to the ambiguous word, the search engine would decide which possibility suits best the empirical data.

A. Source Language Analysis

The parser which is used to analyze and parse the Arabic sentence is described in [8]. It contains a lexicon, and a morphological analyzer. The parser encodes the Arabic grammar rules of irab and the effects of applying these rules to the constituents of sentences. This parser was written in Definite Clause Grammar (DCG) and was developed to be part of a machine translation system. The grammar covers a text from the domain of the agricultural extension documents. After the parsing process of the input Arabic sentence, the Arabic noun phrases are extracted based on the description of the Arabic grammar constituents in the parser. Fig. 2 shows the Arabic grammar constituents as described in [8].

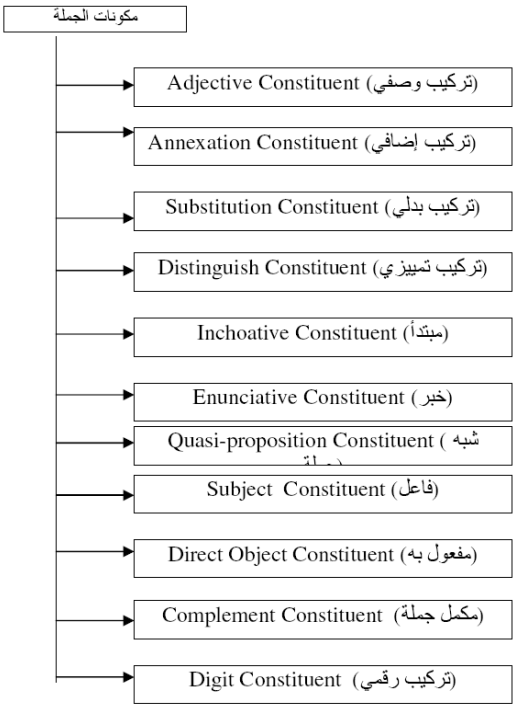

Fig. 2 The Arabic grammar constituents

Noun phrases of sentence are defined in [10], [11] as the maximal syntactic phrases that contain at least one noun and no verb. The Arabic NPs according to this definition are Adjective Constituent, Annexation Constituent, Substitution Constituent, Inchoative Constituent, Enunciative Constituent, Distinguish Constituent, and Digit Constituent.

B. Source to Target Transfer

Syntactic transfer systems rely on mappings between the surface structure of sentences: a collection of tree-to-tree transformations is applied recursively to the analysis tree of the source language in order to construct a target language analysis tree [1]. The tree-to-tree transformation algorithm is a recursive, non- deterministic, top-down process in which one side of the tree-to-tree transfer rules is matched against the input structure, resulting in the structure on the right-hand-side.

In the proposed NP translator, the actual translation occurs in the transfer phase. The Arabic to English transfer involves two steps:

1. Lexical transfer. This maps Arabic lexical units to their English equivalent. It also maps Arabic morphological features to the corresponding set of English features.

2. *Structural transfer*. This maps the Arabic parse tree to the equivalent English syntactic structure.

Lexical translation is performed by a stem-to-stem dictionary. This dictionary is designed to make use of the morphological features which extracted from the morphological analyzer. As described before Arabic word may be translated into many English words because of the script representation of the Arabic words. The Arabic word

with its morphological features such as type(noun, verb, adverb, adjective, …), gender(male, female)and number(single, plural) is used in the translation process to reduce the number of the possible translations of a single Arabic word. Stems are used through the whole translation process in order to reduce data sparseness issues. Stems have a much higher frequency than word tokens, especially in the case of inflected words. Mapping from source to target is one-to-many, meaning that the single word may have more than one translation but in the same time words with different morphological features, become different entries. Taking the morphological features in consideration reduces the ambiguity in the translation process.

The used bilingual dictionary has been extracted from the Golden Al-Wafi® Arabic translator. The Golden Al-Wafi® look up dictionary has a total of 32,653 entries, with between 3 and 4 translations per headword, on average. The morphological features used for Arabic and English come from [8] morphological analyzer. The input of the dictionary look-up is an Arabic word with its morphological features and the output is a set of translation candidates, i.e. strings of English stems.

The structural rules relate other parts and nodes of the two trees to each other. There is a relationship between adjacent lexical units in the Arabic NP. This concerns the preserved order, the relative positioning (precedence), of lexical units in the NP. For example, noun in Arabic precedes its adjective, while in English, adjective precedes the nouns. Consequently, restructuring of the Arabic parse tree is needed to conform to the target English grammar. The structure rules are implemented in Prolog.

C. Target Language Generation

At this point, we are in the Target Language Generation module. Using the morphological features in the previous module reduces the ambiguity in the translation process but still there are ambiguous words. Disambiguation in MT aims to select the correct translation in the target language for an ambiguous item in the source language, based on its context in the translation unit.

In this module a dictionary-graph based WSD approach was investigated which is a combination of dictionary-based WSD [13] and graph-based WSD [19] approaches. It makes use of words which have already been translated as context, implicitly accomplishing basic WSD during the translation process. It uses the Arabic-English dictionary which described before to translate each word in the input Arabic NP. In order to solve the ambiguity in the translation of the Arabic NP , the translation system identify the ambiguous words and the relation between them and use the viterbi search algorithm to find the appropriate translation of the Arabic words to generate the target English NP.

The basic idea is that if we have an ambiguous Arabic word *S* which have two senses S1 and S2. S1 is translated as *T1* and S2 is translated as *T2*. In order to disambiguate an occurrence of *S* in Arabic, we identify the phrase it occurs in and use the viterbi search algorithm to find the appropriate translation of the Arabic word S based on to generate the target English NP.

Viterbi algorithm is a technique which efficiently computes the most likely state sequence. Fig. 3 shows Viterbi algorithm for finding optimal sequence of senses which described in [22].

function VITERBI(*observations* of len *T*,*state-graph*) **returns** *best-path*

num-states ← NUM-OF-STATES(*state-graph*)
Create a path probability matrix *viterbi[num-states+2,T+2]*
viterbi[0,0] ← 1.0
for each time step *t* **from** 1 **to** *T* **do**
  for each state *s* **from** 1 **to** *num-states* **do**

$$viterbi[s,t] \leftarrow \max_{1 \le s' \le num\text{-}states} \left[viterbi[s',t-1] * a_{s',s} \right] * b_s(o_t)$$

$$back\text{-}pointer[s,t] \leftarrow \operatorname{argmax}_{1 \le s' \le num\text{-}states} \left[viterbi[s',t-1] * a_{s',s} \right]$$

Backtrace from highest probability state in final column of *viterbi[]* and return path

Fig. 3 The viterbi algorithm for word sense disambiguation

Given a graph of nodes and weighted edges the algorithm returns the state-path through the graph which assigns maximum likelihood to the observation sequence. $a[s', s]$ is the transition probability from previous state s' to current state s, and $b_s(o_t)$ is the observation likelihood of s given o_t. Note that states 0 and N+1 are non-emitting start and end states. The weights between nodes may be the values of bi-grams or the relatedness between words.

The relation between ambiguous words may be statistical such as statistical measures of association for Ngrams or semantic such as relatedness and similarity. This work combines both statistical and semantic relations between ambiguous words to find the appropriate translation.

1) Statistical Measures of Association for Ngrams

Statistical measures of association judge whether the tokens that make up the Ngram occur together more often than would be expected by chance. If so, then the Ngram may represent a collocation or some other interesting phenomena [5]. A measure that returns a score that can be assigned statistical significance is referred to more precisely as a test of association such as log-likelihood ratio and fisher's exact test. Measures that do not allow for significance to be assigned to their value include the Dice Coefficient and pointwise Mutual Information. When discussing both kinds of techniques we refer to them generically as measures of association.

The different approaches of statistical measures of association were compared in [2] and found that Dice Coefficient is the measure which gave higher accuracy in WSD. In this paper we use the Dice Coefficient as a statistical measure of association combined with semantic relatedness between ambiguous words.

2) Semantic Relatedness and Similarity

Semantic relatedness is used fairly freely, and is sometimes mentioned semantic similarity as well [17]. We should clarify the distinction between these two terms. Two concepts can be related without being similar, so relatedness should be seen as a more general notion than similarity. For example, two concepts may be related because they are antonyms, but they are not likely to be considered similar.

Measures of similarity quantify how much two concepts are alike, based on information contained in an *is–a* hierarchy. *Is–a* relations in WordNet do not cross part of speech boundaries, so WordNet–based similarity measures are

limited to making judgments between noun pairs (e.g., *cat* and *dog*) and verb pairs (e.g., *run* and *walk*). While WordNet includes adjectives and adverbs, these are not organized into *is–a* hierarchies so similarity measures can not be applied. However, concepts can be related in many ways beyond being similar to each other. For example, a *wheel* is a part of a *car*, *night* is the opposite of *day*, *snow* is made up of *water*, a *knife* is used to cut *bread*, and so forth. As such Word-Net provides additional (non–hierarchical) relations such as *has–part*, *is–made–of*, *is–an–attribute–of*, etc. In addition, each concept (or word sense) is described by a short written definition or gloss.

Measures of relatedness are based on these additional sources of information, and as such can be applied to a wider range of concept pairs. For example, they can cross part of speech boundaries and assess the degree to which the verb *murder* and the noun *gun* are related. They can even measure the relatedness of concepts that do not reside in any *is–a* hierarchy, such as the adjectives *violent* and *harmful.*

Three relatedness measures were provided in [18] which are: *hso* [9], *lesk* [4], and *vector* [16].

These relatedness measures were compared in [3] and found that Vector is the measure which gave higher accuracy in WSD. In this paper we use the Dice Coefficient as a statistical measure of association combined with Vector as semantic relatedness between ambiguous words.

We combined Dice Coefficient as a statistical measure of association with the Vector measure of relatedness as following:

$$R(wi,wi\text{-}1)= \mu\ Dice(wi,wi\text{-}1) + \beta\ Vector(wi,wi\text{-}1) + \lambda$$
$$Where:\ \mu + \beta + \lambda=1$$
$$And\ \ 1\geq \mu\geq 0\ ,\ 1\geq \beta\geq 0\ ,\ 1\geq \lambda\geq 0$$

The relation ***R*** is used with viterbi search algorithm to find an appropriate translation of the input Arabic NP.

III. Experiments And Results

These experiments compare the results of statistical, semantic and hybrid statistical-semantic relations between words in the proposed dictionary-graph based WSD approach in an Arabic NP translation system. This allows the comparative evaluation of the different methods to target-language generation, as presented in this paper. A combination of Brown and English Treebank corpus which are available in the Natural Language Toolkit (NLTK) [6] is used as a target language corpus. A corpus of about 1 million words are used and analyzed statistically to get the bi-grams of the words of the corpus. We used the Ngram Statistical Package (NSP) [5]. This package is a set of perl programs that analyze Ngrams in text files. One of these programs takes as input a list of Ngrams with their frequencies and runs a user-selected statistical measure of association to compute a "score" for each Ngram. The Ngrams, along with their scores, are output in descending order of this score. The statistical score computed for each Ngram can be used to decide whether or not there is enough evidence to reject the null hypothesis (that the Ngram is not a collocation) for that Ngram. The statistical measures of association which provided are: dice coefficient, log-likelihood , mutual information, t-score, and the left-fisher test of associativity

To get the relatedness between two words we use WordNet::Similarity [18]. It is a freely available software package that makes it possible to measure the semantic similarity or relatedness between a pair of concepts (or word senses). It provides three measures of relatedness, all of which are based on the lexical database WordNet. These measures are implemented as Perl modules which take as input two concepts, and return a numeric value that represents the degree to which they are related.

Automatic evaluation systems are often criticized for not capturing linguistic subtleties. This is clearly apparent in the field's moving back toward using human evaluation metrics. We conducted a human evaluation of nouns and adjectives realization in a document contained 190 noun phrases. These noun phrases consist of 969 words from them 213 words are ambiguous.

The evaluation was conducted using one bilingual Arabic-English speaker (native Arabic, almost native English). The task is to determine for every ambiguous word that appears in the Arabic input NP whether it is realized or not in the English translation with the correct sense.

In the first experiment we compared the results of bi-gram with Simple Interpolation Smoothing as a baseline with five bi-gram scoring methods, dice, log-likelihood, mutual information, t-score, and the left-fisher test of associativity. The results are presented in Table I

TABLE I: STATISTICAL MEASURES OF ASSOCIATION COMPARED WITH BIGRAM AS A BASELINE

Corpus size 1,171,868 words	
Bi-gram baseline	60.1%
Dice	**63.8%**
Log-likelihood	63.4%
Mutual information	62.0%
T-score	62.9%
Left-Fisher	62.0%

Table I shows that using the statistical measures of association improves the accuracy of WSD by 3.7% and Dice method is the statistical measures of association method which gave the highest accuracy of 63.8%.

In the second experiment we compared the results of semantic relatedness measures hso, lesk and vector. The results are presented in Table II.

TABLE II: SEMANTIC RELATEDNESS MEASURES

HSO	38.0%
Lesk	51.64%
Vector	**53.52%**

Table II shows that vector method is the semantic relatedness measure with the highest accuracy of 53.52% but it did not improve the accuracy of WSD.

In the third experiment we combined the statistical and semantic measures together. The results are presented in Table III.

TABLE III: THE HYBRID SEMANTIC-STATISTICAL MEASURES

Dice-HSO	62.91%

Dice-Lesk	66.19%
Dice-Vector	**68.08%**

Table III shows that the hybrid semantic-statistical method improved the accuracy of WSD by 4.28% and Dice-Vector combination is the hybrid measure which gave the highest accuracy of 68.08%.

The improvement in WSD strongly affects the accuracy of MT. In the baseline MT there was 170 ambiguous noun phrases. As shown in table IV when applying WSD using Dice the number of ambiguous noun phrases decreased to be 39 phrases with translation accuracy of about 69%. WSD with Vector decreases the number of ambiguous noun phrases to be 56 phrases with translation accuracy of 60%. WSD with hybrid Dice-Vector gave the highest improvement in MT process it decreased the number of ambiguous noun phrases decreased to be 32 phrase with translation accuracy of about 73%

TABLE IV: THE ACCURACY OF THE TRANSLATION SYSTEM

	No. of Ambiguous NPs	**Ambiguity**	**Accuracy**
Baseline system	170	89.5%	10.5%
Dice	39	31.05%	68.95%
Vector	56	40 %	60%
Dice-Vector	**32**	**27.37%**	**72.63%**

IV. CONCLUSION AND FUTURE WORK

Translation systems built only on a conventional dictionary and monolingual corpora might be useful for applications such as document classification or multilingual information retrieval. The general design of the proposed approach allows for translation without an extensive rule-set or a parallel corpus.

This paper presents an approach to improve the accuracy of translation process in a machine translation system that translates Arabic noun phrase into English using a dictionary-graph based WSD approach. We presented a hybrid semantic-statistical method based on computing words relatedness and a statistical measure of association to get the relation between ambiguous words. This relation was used with viterbi search algorithm to find the appropriate translation of the Arabic NP. This paper shows that using hybrid methods could achieve an improvement in WSD compared with statistical and semantic methods and also improve the accuracy of the translation system as well.

Manual evaluation of 213 ambiguous words in 190 NPs was accomplished for all scoring methods.We fined that the accuracy of all semantic measures increased when combined with statistical measured. The final experiment with showed that dice-vector combination is the hybrid measure which gave the highest accuracy. It improved the WSD accuracy by 4.28% and decreased the number of ambiguous noun phrases in the translation process from 170 noun phrases to be 32 phrases with translation accuracy of 73%.

When building a hybrid system, using rules and statistics, it is important to keep the number of rules limited, to ensure that the system can be transferred to other language pairs, without spending large amounts of time on rule-writing.

To improve the translation process in future work we need larger corpus. Using the morphological features with the translated words may also improve the output.

REFERENCES

[1] A. Abd El-Monem, "Machine Translation of Noun Phrases: from English to Arabic," M. Sc. Thesis, Faculty of Engineering, Cairo University, Giza, Egypt, 2000.

[2] O. M. Ali, M. GadAlla, and M. S. Abdelwahab, "Word Sense Disambiguation in Machine Translation using Monolingual Corpus," Proceedings of The Eighth Conference on Language Engineering, Ain Shams University, Cairo, Egypt, 2008, pp. 141-151.

[3] O. M. Ali, M. GadAlla, and M. S. Abdelwahab, "Improving Word Sense Disambiguation in Machine Translation using Semantic Relatedness and Statistical Measures of Association," Proceedings of The Forth Intenational Conference on Intelligent Computing & Information Systems (ICICIS-2009), Faculty of Computer and Information Science Ain Shams University, Cairo, Egypt, 2009, pp. 885-859.

[4] S. Banerjee, and T. Pederse, "Extended gloss overlaps as a measure of semantic relatedness," In Proceedingsof the Eighteenth International Joint Conference on Artificial Intelligence, 2003, pp. 805–810

[5] S. Banerjee, and T. Pedersen, "The Design, Implementation and Use of the Ngram Statistics Package," Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, 2003 ,February, Mexico City.

[6] Bird, Steven and E. Loper, "Natural Language Toolkit," 2006. http://nltk.sourceforge.net/

[7] Y. S. Chan, H. T. Ng, and D. Chiang, "Word Sense Disambiguation Improves Statistical Machine Translation," Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, June 2007, p.p 33–40.

[8] A. Farouk, "Developing an Arabic Parser in a Multilingual Machine Translation System," M. Sc. Thesis, Computer and Information Sciences, Institute of Statistical Studies and Researches, Cairo University, Giza, Egypt.1999.

[9] G. Hirst, and D. St-Onge, "Lexical chains as representations of context for the detection and correction of malapropisms," in: C. Fellbaum (Ed.), WordNet: An electronic lexical database, MIT Press, 1998, pp. 305–332.

[10] P. Koehn, "Noun Phrase Translation," Ph.D. Thesis, University of Southern California, 2003.

[11] P. Koehn, and K. Knight, "Feature-rich statistical translation of noun phrases," In Proc. of the 41st Annual Meeting of the ACL, Sapporo, Japan, 2003.

[12] S. Legrand, JGR Pulido, "A Hybrid Approach to Word Sense Disambiguation: Neural Clustering with Class Labeling," In: Knowledge Discovery and Ontologies (KDO-2004) workshop, 15th European Conference on Machine Learning (ECML) and 8th European Conference on Principles and Practice of Knowledge Discovery in Databases , Pisa, Italy, September 24, 2004.

[13] C. D. Manning, and H. Schutze, "Foundations of Statistical Natural Language Processing," the MIT press, 1999, ch. 7.

[14] T. M. Miangah, and A. D. Khalafi, "Statistical analysis of target language corpus for word sense disambiguation in a machine translation system," In 7th EAMT Workshop, "Broadening horizons of machine translation and its applications", Malta, 26-27 April 2004, pp. 129-137.

[15] M. Mihăltz, A. Towards, "Hybrid Approach To. Word-Sense Disambiguation In Machine Translation," In Proceedings of International Workshop, "Modern Approaches in Translation Technologies Borovets", Bulgaria, 2005.

[16] S. Patwardhan, "Incorporating dictionary and corpus information into a context vector measure of semantic relatedness," M. Sc. Thesis, University of Minnesota, Duluth, 2003.

[17] T. Pedersen, S. Banerjee, and S. Patwardhan, "Maximizing Semantic Relatedness to Perform Word Sense Disambiguation," University of

Minnesota Supercomputing Institute Research Report UMSI 2005/25, March.
[18] T. Pedersen, and S. Patwardhan, and J. Michelizzi, "WordNet::Similarity - Measuring the Relatedness of Concepts," In: Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-2004) San Jose, CA. July, 2004, pp. 1024-1025.
[19] R. Sinha, and R. Mihalcea, "Unsupervised Graph-based Word Sense Disambiguation," book chapter in ``Current Issues in Linguistic Theory: Recent Advances in Natural Language Processing", Editors Nicolas Nicolov and Ruslan Mitkov, John Benjamins Publishers, 2009.
[20] G. Thurmair, "Improving Machine Translation Quality," In Proceedings of the Tenth Machine Translation Summit. Phuket, Thailand, 2005.
[21] A. Franz, K. Horiguchi, L. Duan, D. Ecker, E. Koontz, and K. Uchida, "An Integrated Architecture for Example-Based Machine Translation.Spoken Language Technology", Sony US Research Labs,USA, 2000.
[22] D. Jurafsky and J. H. Martin, "Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition," Prentice Hall, 2000, ch. 5.